What's an Expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis.

I. WORSLEY
Stanford University, U.S.A.
AND
II.BLIKSTEIN
Stanford University, U.S.A.

Assessing student learning across a variety of environments and tasks continues to be a crucial educational concern. This task is of particular difficulty in non-traditional learning environments where students endeavor to design their own projects and engage in a hands-on educational experience. In order to improve our ability to recognize learning in these constructionist environments, this paper reports on an exploratory analysis of learning through multiple modalities: speech, sentiment and drawing. A rich set of features is automatically extracted from the data and used to identify emergent markers of expertise. Some of the most prominent markers of expertise include: user certainty, the ability to describe things efficiently and a disinclination to use unnecessary descriptors or qualifiers. Experts also displayed better organization and used less detail in their drawings. While many of these are things one would expect of an expert, there were areas in which experts looked very similar to novices. To explain this we report on learning theories that can reconcile these seemingly odd findings, and expound on how these domain-independent markers can be useful for identifying student learning over a series of activities.

Key Words and Phrases: Learning Analytics, Multi-modal, Assessment, Speech

1. INTRODUCTION

The call to improve assessment of student learning is being raised from various fronts. National education policy mandates that schools demonstrate student advancement on a regular basis. At the same time, corporations and public institutions call for students to learn 21st century skills: creativity, collaboration and innovation. Educators, who scramble to satisfy these competing demands, find themselves at an irreconcilable crux. One solution may lie in the development of automatic natural assessment tools. Such tools can provide the automaticity needed to allow testing to be more open-ended and also offer innovative teachers new ways for assessing how their students are learning during hands-on, student-designed learning. With this in mind, our primary research question is: How can we use informal student speech and drawings to decipher meaningful "markers of expertise" (Blikstein 2011) in an automated and natural fashion?

2. PRIOR WORK

This research builds on a growing tradition in artificial intelligence in education that uses various techniques to uncover correlations between student artifacts and efficacious learning. Previous work includes a variety of examples from intelligent tutoring systems that leverage: discourse analysis (Litman et al 2009, Forbes-Riley et al 2009), content word extraction (Chi et al 2010, Litman et al 2009), uncertainty detection (Liscombe et al 2005), sentiment analysis (Craig et al 2008, D'Mello et al 2008, Conati 2009), linguistic

analysis, prosodic and spectral analysis, and multi-modal analysis (Litman et al 2009, Forbes-Riley and Litman 2010). Other examples from the education context include automatic essay grading (Chen et al 2010, Rus et al 2009) and educational robots. The present study also extends our previous work (Blikstein and Worsley 2011) and explores the salience of the aforementioned analysis techniques in characterizing open-ended learning.

3. DATA

The data for this study comes from interviews with 15 students from a tier-1 research university. Of the 15 students, 8 were women, 7 were men; 7 were from technical majors, 3 were undergraduates, and 12 graduate students. There were 3 novices, 9 intermediates, and 3 experts and each interview took approximately 30 minutes. Participants were asked to draw and think aloud about how to build various electronic and mechanical devices. The questions were posed in a semi-structured clinical interview format. Question 1, the control question, asked the student to construct a temperature control system, while question 2 challenged the student to design a device to automatically separate, glass, paper, plastic and metal. Student speech was transcribed by graduate and undergraduate students. Prior to the interviews, the subjects were labeled as being experts, intermediates or novices in engineering and robotics. This classification was based on previous formal technical training either through a degree program or through a lab course on physical computing. This classification is in accordance with theory that suggests that experts are those that have had extended time practicing their skill. The data consisted of audio files, transcriptions of the interviews, and digitized drawings that the students produced during the interview.

4. DATA ANALYSIS

In accordance with previous literature, this study utilized the following techniques for feature extraction: crowd-sourcing using Mechanical Turk to determine human ratings of each transcript; prosodic analysis - pitch, intensity and duration – and spectral analysis – the first three formants - using the Praat software (Boersma and Weenick, 2010); linguistic analysis - pauses, filled pauses, restarts – using the Python Natural Language Toolkit; sentiment analysis using the Linguistic Inquiry and Word Count (LIWC) and the Harvard Inquirer; content word analysis, using web-mined lexicons from chemistry, mathematics, computer science, material science and general science; dependency parsing using the Stanford Parser (Klein and Manning, 2003); n-gram analysis; and human coded drawing analysis based on the work of Song and Agogino 2004, Shah et al. 2003 and Anderson 2000.

The data was analyzed using expectation maximization (EM) with an intra-cluster Euclidean distance objective function. Before running EM, each feature value was modified to have unit variance and zero mean. Additionally, t-tests were performed to check statistical significance.

5. RESULTS

The complete analysis involved nearly 200 features, excluding n-grams. For the sake of brevity, we will only report on a subset of features. From the drawing analysis data in Table 1, we see little variation across the classes. Moreover, the only statistically significant class differences exist between novices and non-novices. These differences are observed for ‘space used’ and dimensions.

Table 1 - A comparison of the average drawing features scores across expertise types. Text Annotation ranges from 0 to 1, while, Space Used, Is 3-D and dimensions range from 1 to 3. Finally, the remaining scores were rated on a 10 point scale.

Class	Text Annotation	Space Used	Abstraction	Is 3-D	Detail	Organization	Dimensions
Novice	1.00	2.47	5.97	1.64	5.83	5.94	2.11
Intermediate	0.78	1.85	4.60	1.63	4.31	5.46	2.05
Expert	1.00	1.92	6.27	1.10	4.85	8.25	1.55

Similar class-based statistics are reported for significant linguistic, prosodic and sentiment features. Table 2 presents the linguistic and prosodic results, while Table 3 presents the sentiment analysis results.

Table 2 – The normalized average duration, pitch, intensity and number of disfluencies, among the different classes.

Expertise	Duration	Pitch	Disfluencies	Intensity
Novice	1.16	0.7	1.06	-0.39
Intermediate	-0.2	-0.09	-0.52	0.02
Expert	-0.56	-0.42	0.5	0.32

To provide additional clarification about the linguistic and prosodic data trends, consider that the average duration of a novice answer was nearly twice as long as that of an expert.

Table 3 – Average normalized word count for various sentiment words from LIWC and the Harvard Inquirer

Class	Positive	Strong	Weak	Understate	Quality	Quantity	Certainty
Novice	0.063	0.058	0.032	0.053	0.020	0.065	0.014
Intermediate	0.041	0.068	0.033	0.039	0.028	0.068	0.022
Expert	0.039	0.077	0.024	0.036	0.012	0.079	0.022

Finally, we present the centroids that were obtained from doing clustering analysis with both sentiment and speech features.

Table 4 - EM cluster centroid values for the features included in a combined sentiment and speech analysis. Values have been normalized to unit variance and zero mean. Duration is in seconds, while the other values are in words per transcript.

	Novice	Intermediate	Expert
Duration (s)	0.97	-0.30	-1.23
Filled Pauses	-0.13	-0.51	1.78
LIWC Neg	0.32	-0.52	1.3
SureLw	-0.68	0.26	0.65
Quality	-0.55	0.62	-1.1

6. DISCUSSION

Of particular interest to this study is the presence of several features that accurately predict expertise based on certainty. Though not presented at length, preliminary analysis of this data involved extracting n-grams from each transcript and looking for patterns across the different classes. Not surprisingly, n-grams that indicated uncertainty, eg. “don’t know”, “well, you know” were more common among novices than among non-novices. These initial results confirm a theory previously presented by Beck, in Bruer (1993) which indicates that increasing expertise tends to increase student self-confidence. These results were further corroborated in our later analysis through the certainty (SureLw) and understatement (Undrst) features. Certainty was much more common among experts, while understatements were more frequently employed by novices. Furthermore, we saw subtle leanings towards certainty through the “strong” and “weak” features, which were more prevalent among experts and novice, respectively.

The observed results concerning the decrease in duration for intermediate and expert participants, as compared to novices, also suggests that more advanced users are more certain in their approach. However, the decrease in duration is also in accord with work by Anderson and Schunn’s ACT-R theory, which describes how experts have greater facility in accessing the necessary declarative and procedural skills needed to solve complex problems, simply due to their increased exposure to them. More specifically, Anderson and Schunn (2000) completed a similar study in which they observed a substantial decrease in time needed to complete geometry proofs as students spent more time working on them.

Somewhat unexpected was the lack of meaningful results from the drawing analysis and content word analysis. The initial hypothesis for the drawing analysis assumed that more expert individuals would be capable of providing superior drawings of the system because of an improved mental representation of the required components (Anderson 2000). Instead we found little to no correlation between our features and expertise. Even in the case of our organization metric which showed a statistically significant difference between classes, the correlation coefficient was 0.13. We attribute some of this ambiguity to the drawings having a different audience for different research participants. Certain participants viewed the drawings as artifacts that they were making for the researchers, whereas others viewed the drawing space as a place for them to take notes, and simply get their thoughts on paper.

Similarly, the content word analysis failed to provide meaningful features for distinguishing experts from novices. While this may suggest that the task was not sufficiently difficult, a more likely explanation may be related to the informal nature of the interaction. According to Brown and Spang (2008) the language of science and mathematics are decidedly different from the language of everyday conversation. Because of this, it is unlikely that students will employ noticeably different levels of science and mathematics terminology in informal settings. Additionally, the nature of this open-ended design space is that people will bring previous knowledge from a variety of backgrounds and use that to solve problems. As such, it could be perfectly conceivable for a computer scientist, chemical engineer and mechanical engineer to all come up with expert solutions to a problem using completely different nomenclature.

Taken together, these results provide additional validation for the need to develop novel assessment techniques that leverage natural student artifacts: speech and drawings.

7. CONCLUSION

This study has explored a set of domain-independent markers of expertise that can allow educators and researchers to recognize student learning through analyzing student

speech, and, to a lesser extent, drawings. Using speech as a form of assessment certainly presents some challenges, but has the potential to introduce innovative ways for understanding and predicting learning in open-ended learning environments. This ability to assess non-traditional learning should help open the door to more widespread adoption of experiential learning practices, and an associated increase in 21st century competencies. Thus far our work has been exploratory. We performed in-depth analysis on a small sample size in order to better inform the types of features that we need to be looking for in future work. This initial work points to user uncertainty, as perceived through various modalities, as an influential indicator of student development. In future work we plan to further validate our findings through larger scale, longitudinal studies in constructionist learning environments.

REFERENCES

Harvard Inquirer. http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm

Linguistic Inquiry and Word Count http://liwc.net/liwcdescription.php

ANDERSON, J.R. 2000. Cognitive Psychology and Its Implications, Fifth Edition. Worth Publishing.

ANDERSON, J.R. & SCHUNN, C.D. 2000. Implications of the ACT-R Learning Theory: No Magic Bullets in R. Glaser (Ed.), *Advances in instructional psychology (Vol. 5)*. Mahwah, NJ.

BLIKSTEIN, P. 2011. Using learning analytics to assess students' behavior in open-ended programming tasks. *Paper presented at the I Learning Analytics and Knowledge Conference*, Banff, Canada.

BLIKSTEIN, P. & WORSLEY, M. 2011. Learning Analytics: Assessing Constructionist Learning Using Machine Learning. *Paper presented at the American Educational Research Association Annual Meeting*, New Orleans, USA.

BOERSMA, P., AND WEENINK, D. 2010. Praat: doing phonetics by computer [Computer program]. Version 5.2.03, http://www.praat.org/. Design Studies 19: 431-453

BROWN, B. A., AND SPANG, E. 2008, Double talk: Synthesizing everyday and science language in the classroom. *Science Education*, 92: 708–732.

BRUER, J.T. 1993. Schools for Thought: A science of learning in the classroom. MIT Press.

CHEN, Y., LIU, C., LEE, C., AND CHANG, T. 2010, "An Unsupervised Automated Essay Scoring System," *Intelligent Systems, IEEE* , vol.25, no.5, pp.61-67, Sept.-Oct. 2010

CHI, M., VANLEHN, K., LITMAN, D., AND JORDAN, P. 2010. Inducing Effective Pedagogical Strategies Using Learning Context Features. In: *Proc. of the 18th Int. Conference on User Modeling*.

CONATI, C. AND MACLAREN, H. 2009. Empirically building and evaluating a probabilistic model of user affect. U*ser Modeling and User-Adapted Interaction,* August, 2009 267-303.

CRAIG, S. D., D'MELLO,S., WITHERSPOON, A. AND GRAESSER, A. 2008. 'Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive-affective states during learning', *Cognition & Emotion*, 22: 5, 777 — 788.

D'MELLO, S. K., CRAIG, S. D., WITHERSPOON, A., MCDANIEL, B., AND GRAESSER, A. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction* 18, 1-2 (Feb. 2008), 45-80.

FORBES-RILEY, K., AND LITMAN, D. 2010. Metacognition and Learning in Spoken Dialogue Computer Tutoring. *Proceedings 10th International Conference on Intelligent Tutoring Systems (ITS),* Pittsburgh, PA.

FORBES-RILEY, K., ROTARU, M., AND LITMAN, J. 2009. The Relative Impact of Student Affect on Performance Models in a Spoken Dialogue Tutoring System. *User Modeling and User-Adapted Interaction (Special Issue on Affective Modeling and Adaptation)*, 18(1-2), February, 11-43.

KLEIN, D AND MANNING, C. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

LISCOMBE, J., HIRSCHBERG, J., AND VENDITTI, J. 2005. Detecting Certainness in Spoken Tutorial Dialogues. In *Proceedings of Interspeech 2005—Eurospeech*, Lisbon, Portugal.

LITMAN, D., MOORE, J., DZIKOVSKA, M., AND FARROW. E. 2009. Using Natural Language Processing to Analyze Tutorial Dialogue Corpora Across Domains and Modalities. *Proceedings 14th International Conference on Artificial Intelligence in Education (AIED)*, Brighton, UK, July.

LITMAN, D., AND FORBES-RILEY, K. 2009. Spoken Tutorial Dialogue and the Feeling of Another's Knowing. *Proceedings 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, London, UK, September.

RUS, V., LINTEAN, M.,, AND AZEVEDO, R.. 2009. Automatic Detection of Student Mental Models During Prior Knowledge Activation in MetaTutor. In *Proceedings of the 2nd International Conference on Educational Data Mining* (Jul. 1-3, 2009). Pages 161-170

SHAH ,J., VARGAS-HERNANDEZ, N., SMITH, S.M. (2003) Metrics for measuring ideation effectiveness. *Design Studies* 24: 111-134

SONG, S., AGOGINO, A.M. 2004 Insights on Designers' Sketching Activities in Product Design Teams. 2004